DOCUMENT RESUME

ED 252 563 TM 850 032

AUTHOR Herman, Joan
TITLE A Practical Approach to Local Test Development. Resource Paper No. 6. Research into Practice Project.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Nov 84
GRANT NIE-G-84-0112-P4
NOTE 63p.
PUB TYPE Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Elementary Secondary Education; Instructional Development; Models; Rating Scales; Student Evaluation; *Teacher Made Tests; *Test Construction; Test Format; *Test Items; Test Manuals; Test Validity
IDENTIFIERS Curriculum Related Testing; *Domain Referenced Tests; *Test Specifications

ABSTRACT

This resource paper is a guide for planning and developing instructionally relevant tests of student learning at the classroom, building, or district level. It is based on a model of instruction and testing which systematically uses assessment information to support and facilitate instructional improvement. Course goals and objectives are first translated into domain specifications which are used to link testing and instruction. A domain specification contains six components: (1) domain description; (2) content limits; (3) distractor limits and response-criteria; (4) item format; (5) student directions; and (6) sample items. Test items are then developed to match domain specifications. Standard item construction rules are given for both constructed responses (essay, short answer, completion) and selected responses true-false, matching, multiple choice). Items are then judged for their match to the six domain components plus linguistic and thinking complexity using the Item Rating Scales. When all eight categories have been individually scored on scales from 0 to 10, an overall rating is calculated. Guidelines are given for interpreting overall item rating for acceptance, rejection, or revision. Appendices contain: (1) sample domain-referenced test specifications; (2) item difficulty levels; and (3) materials used in the item rating process. (BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED252563

DELIVERABLE - NOVEMBER 1984

RESEARCH INTO PRACTICE PROJECT

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
[ ] Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Joan L. Herman
Project Director

Resource Paper No. 6, 1984

A Practical Approach to Local Test Development

Grant Number
NIE-G-0112-P4

TM 850 032

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

A PRACTICAL APPROACH TO LOCAL TEST DEVELOPMENT

James Burry
Joan Herman
Eva L. Baker

Resource Paper No. 6
1984

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was supported pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by the National Institute of Education should be inferred.

TABLE OF CONTENTS

	PAGE
INTRODUCTION	1
Potential Users of the Guide	1
Approach to Testing	2
Structure of the Guide	5
THE COMPONENTS OF DOMAIN SPECIFICATIONS	5
Overview to Components	5
Analyzing Intentions & Expectations	5
Developing the Domain Specification	6
Domain Description	9
Content Limits	13
Distractor Limits/Response Criteria	13
Format	16
Directions	16
Sample Item	17
Summary	18
Sample Domain Specification	19
ITEM CONSTRUCTION RULES	20
Constructed Responses	21
Selected Responses	23
THE ITEM RATING SCALE	29
Background	29
Using the IRS	29
Overall Item Rating	35
Interpreting an Item's Overall Rating	36
USING THE GUIDE	38
CONCLUSION	40
REFERENCES	41
APPENDIX A: SAMPLE DOMAIN SPECIFICATIONS	42
APPENDIX B: DIFFICULTY LEVELS	50
APPENDIX C: RATING MATERIALS	52

INTRODUCTION

Potential Users of the Guide

This resource paper offers a guide for planning and developing instructionally relevant tests of student learning. The planning and development approach we describe responds to findings from several years of CSE research on the uses of tests and the broader evaluation systems of which they are often a part. The heart of these findings is that school practitioners needs tests which match their curriculum, which are useful for instructional planning, and which are fair and valid for evaluation.

For example, at the classroom level, teachers rely to a great extent on the results of tests that they themselves develop in large part because their tests are sensitive to their instructional intentions and are viewed as most appropriate for their students in both content and format. They want tests that reflect their instruction and that provide information they can use to monitor student learning (Dorr-Bremme, 1983).

At the building level, principals too want information that can be used to judge their schools' progress. Like teachers, principals want tests which match their actual school programs but are hesitant to use the results of teacher developed tests (Burry et al, 1982). Perhaps, like some others, they have reservations about the quality of such tests.

Those at the district level also echo concern for the instructional relevance of testing programs. Complaints are often made that the standardized, norm-referenced tests which are frequently administered do not match up with districts' instructional offerings and intentions (O'Shea, 1981) and are inappropriate for accountability purposes. In response, more and more districts are developing their own tests to match their curricular continuum (Burry et al, 1981).

The test development process described in this guide reflects the need for tests which match curriculum and instruction. The specification of curricular and instructional intentions, in fact, is the core of the development process. Because these intentions guide item development, the validity and usefulness of the testing is increased.

Because instructional intentions can be defined at the class, school, or district levels, the test development process described in the guide is useful in creating valid and useful tests for all these levels. The guide is thus appropriate for a variety of users:

- groups of teachers and their principals can use the guide to develop tests that reflect classroom/building needs;
- district testing specialists can use the guide where there is a need for tests, in addition to those developed for teacher use, that reflect district progress;
- groups of teachers and district staff can use the guide and work together to meet both kinds of needs.

Approach to Testing

Is this guide any different from other materials purporting to be of value in local test development efforts? We think it is, because it reflects a concern for fairness and utility in testing. It leads to tests that match curricular intentions and have relevance for instructional decision making. It emphasizes the need to integrate the acts of teaching and testing. What is taught provides the basis for what will be tested, and the results of the testing can then feed back to the ongoing business of teaching.

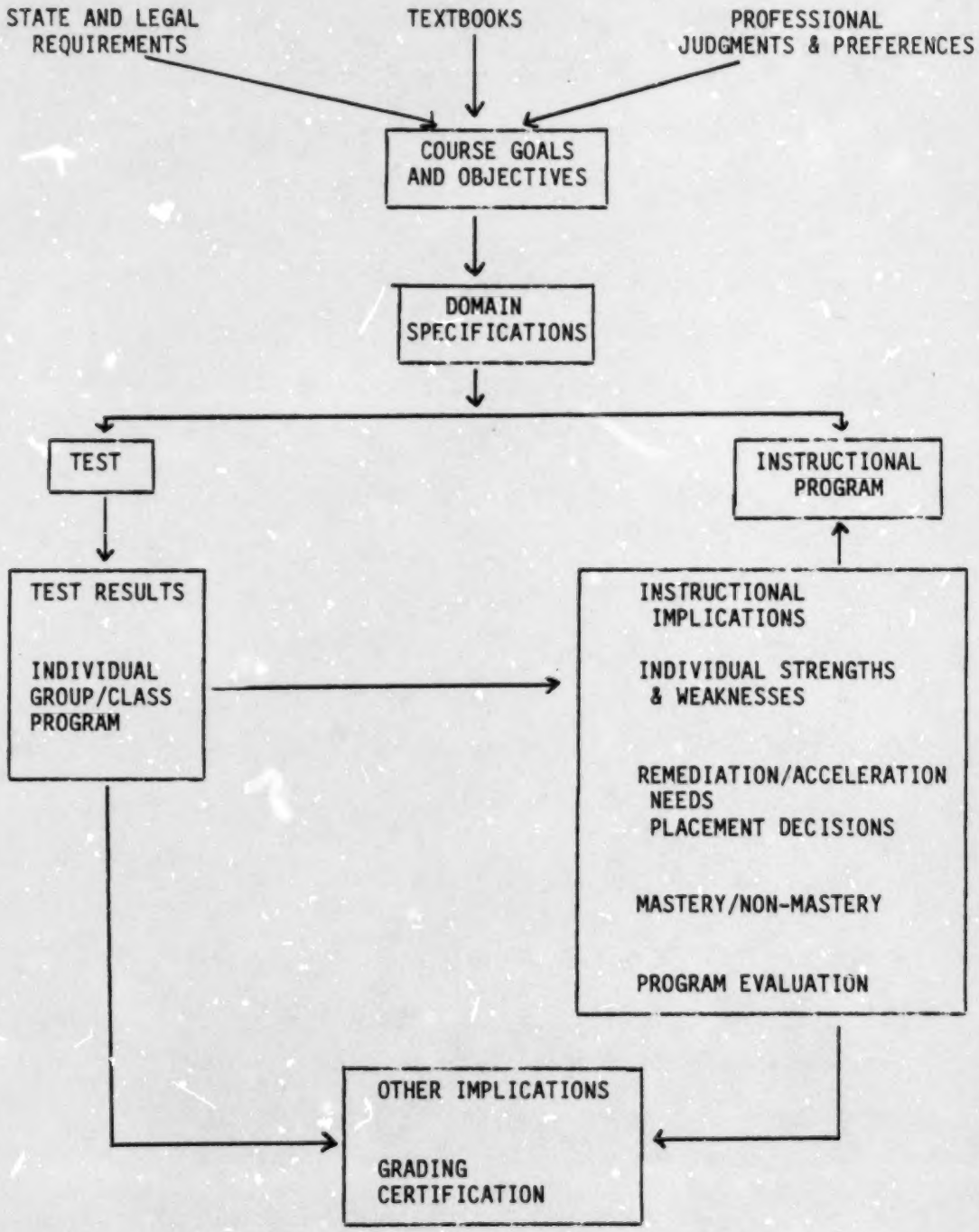
The idea is not, as some have suggested, that tests ought to drive the curriculum, nor that teachers, strictly speaking, ought to "teach to the test." Rather, both testing and instruction ought to reflect significant, agreed upon curriculum goals and objectives. Tests should measure important class, school, and district objectives, and classroom instruction should provide students with an opportunity to attain those objectives. The model, as displayed in Figure 1, is a simple one.

Figure 1 displays a model of instruction and testing which systematically uses assessment information to support and facilitate instructional improvement. As the figure implies, state, district, legal and other requirements, available tests and other instructional materials, and professional judgments are synthesized to arrive at goals and objectives. These goals and objectives then serve as the guidepost for designing instruction and tests. Because both testing and the instructional program mirror the same goals and objectives, test results can be used to identify areas where individuals may need more help, where additional class instruction is needed, and where the instructional program (the next time around) can be strengthened and improved. Because they match what schools are trying to accomplish, the results also provide a fair and valid measure of effectiveness.

You'll notice that Figure 1 includes an additional element, labeled "domain specifications." These specifications clarify the nature of the goals and objectives that are to be taught and provide a conceptual map that can guide both testing and instruction. They likewise provide a public and open model of exactly what is expected al all -- a clear statement of the knowledge, content, skills, and procedures that teachers

Figure 1*

DOMAIN REFERENCED VIEW OF INSTRUCTION AND TESTING



* Taken from Herman, J.L., Testing and Instructional Improvement: An Integrated Test Development Process.

intend to teach and that students are expected to learn. Domain specifications, as described later, are the most arduous part of the test development process. They are also the critical link which enables the integration of testing and instruction and helps to assure that tests are sensitive to a school's instructional program, are targeted on meaningful skills, and that the entire testing process is fair and useful.

Structure of the Guide

The guide is set out as follows:

We begin with a description of domain specifications and exemplify each of the major components included in the specification. This section develops ongoing illustrations of each domain component and concludes by offering a complete sample domain-referenced test specification. Others are provided in Appendix A.

The next section of the guide offers some generally-held principles of item construction for each of the major item forms in the constructed response and the selected response modes.

The final section offers procedures for ensuring that items written to assess a given domain do indeed match their specifications. We provide a scale for this purpose, along with procedures for using the scale to judge an item's fit with each of the elements in the domain specification, for interpreting the meaning of the final rating of fit the item receives, and for deciding what that rating implies for modifications in the item or its specification.

THE COMPONENTS OF DOMAIN SPECIFICATIONS

Overview to Components

A domain specification includes six major components as follows:

First, the domain description focuses on what's expected of the

student in a particular area.

Second, the content limits set the range of content that can be used to write test items. This step has an option for developing selected response items or constructed response items.

A selected response item presents the student with a question or problem and alternative answers to the question. The student's job is to pick the correct answer. A constructed response item, on the other hand, asks the student to create an answer for a question or problem.

Third, the distractor limits describe the wrong answers that may be used as alternatives for selected response items. The response criteria, which is the counterpart to distractor limits, provide the rules for judging a student's constructed response.

Fourth, the format describes the item presentation form.

Fifth, the directions tell the student what he or she is supposed to do in answering the questions.

And sixth, a sample item, reflecting the rules in 1 to 5 above, is provided.

Each of these components is clarified in the following sections.

Analyzing Intentions and Expectations

Domain specifications begin by considering what is to be taught and assessed. The nucleus of a domain specification might begin by stating the principal outcome expected of students: For example:

° Writing a paragraph

If writing a paragraph were to be the subject of a full domain description, what would be some of the skills we might expect of students as they write their paragraphs? Perhaps they would be:

° stating a main idea

° offering supporting details

° using complete sentences to form the paragraph

° using correct spelling, punctuation, and grammar

Let's take a look at the nucleus of another domain specification. Perhaps we want students to be able to

° Identify triangles

and after instruction we would want students to be able to select as triangles from among various geometric shapes only those which have:

° three sides

° straight sides

° enclosed shape

The instructional implications of such a domain specification should be obvious; that is, how the specification can be used to identify critical features of both instruction and testing. For example, we have said that the defining features of triangles are three sides, straight sides, and a closed shape. But these features represent only a preliminary definition. Even in this simple example, a number of instructional questions can be raised about the kinds of discriminations, within the broad domain, that we want students to be able to make. Do we want them to be able to identify the generic triangle shape only? To be able to identify isosceles triangles? equilateral triangles? right triangles? Some of these?

All of these? Why? Do we wish to set conditions or barriers that we want students to be able to cope with as they go about identifying triangles, such as triangles standing on their vertices as opposed to their bases? If so, why?

We raise these issues to underscore, once again, the notion that domain specifications attempt to integrate the acts of instruction and assessment. They do this by precisely specifying instructional intentions, expected student outcomes, and accurate measures of these outcomes. The nature of these intentions and expectations guide the development of the six major components of the domain specification which we alluded to earlier, and which we'll now begin to look at more closely.

Developing the Domain Specification

The first thing to do before writing a domain specification is to set its broad focus. First, what is the subject matter that is to be tested -- math; English mechanics; English composition? Second, what is the intended grade level of the instruction and the assessment, and how might this level affect the readability and intended difficulty of the items that are to be developed on the basis of the domain specifications? Third, what kind of items will be used -- selected response; constructed response? Fourth, at what cognitive level do we want the students to operate in demonstrating their knowledge during assessment -- knowledge; comprehension; application; analysis; synthesis; evaluation? (See Appendix B for a description of each of these levels.)

The answers to these questions help to keep the domain specification properly focused as we begin to write up its components. If we plan to develop a specification for selected response items, our specification will

contain: Domain description; content limits; distractor limits; format, directions, and sample item. If we plan to develop a specification for constructed response items, our specification will differ from the above description only by replacing distractor limits with response criteria.

In the next section, we will describe each of these domain specification features and give examples as we go along.

Domain Description

The domain description provides a broad but operational definition of the behavior expected of the students in a particular content area. This description may consist of an objective or an explanation of a task and/or its components. It may include performance conditions, although these conditions will be specified in greater detail later in the specification.

Here are some examples of domain descriptions:

- Math -- identifying shapes as triangles
- English mechanics -- applying the rules of capitalization
- English composition -- writing a legible, well-organized, and grammatically correct paragraph in which a position is taken and supported.

This description should give a specific picture of the domain of interest to the person who will use the domain specifications as a blueprint for developing test items.

Content Limits

Content limits describe the ballpark from which items can be written. If the item does not fit in the ballpark, then it is not assessing what we want it to assess. Therefore, the content limits must provide a careful description of the range of eligible content from which test items may be

written. This description may include rules for creating questions, rules for generating prompts, cues, or additional materials, such as pictures, graphs, reading selections.

Note that here we are talking about the "question" or task part of the item; that is, the stem. Other parts of the item, such as its distractors or its scoring criteria, are specified later in the process.

The nature of the content limits will vary depending on whether we are writing a domain description for selected response items or for constructed response items.

For Selected Response Items: A selected response item asks the student to choose an answer from a number of given alternatives such as true-false, matching, multiple choice. Content limits for selected response items, therefore, need to define and restrict the characteristics of the item stem and any additional material included in the presentation of the question or problem.

Here are two examples of content limits, building on the first two examples we began with in the domain description section, for selected response items:

- Math -- the student will be asked to select the triangle from among four shapes, only one of which is a triangle. Permissible shapes will include 4 or more sided figures which are linear; 3 sided figures in which one side is curvilinear and circles. Triangles included in the test will reflect the following: equilateral, isosceles, obtuse, and acute.

- English mechanics -- the student will be presented with a sentence and asked to select the word that is improperly capitalized. The sentence will contain at least four capitalized words, one of which is improperly capitalized. The following rules will be used in determining correct and incorrect capitalization.

Let's take a look at these content limits for a moment. In the math content limit, in this example, we are making a rule that only four shapes can be presented, and only one of them can be a triangle. This means that an item containing three shapes, five shapes, or more than one triangle would not meet the specified content limits. More importantly, we are also describing the kinds of shapes that can be used in the assessment.

In the content limits described for the Engish mechanics domain, we are specifying that the sentence contain at least four capitalized words, and only one of them can be improperly capitalized. An item with three capitalized words would not match the content limits as described in the example; a sentence with six capitalized words would. A sentence with more than one improperly capitalized word would also fail to match the described content limits. And more substantively, a capitalized word which did not exemplify one of the specified rules would violate the specification and be an unfair measure of instruction.

For constructed response items. Constructed response items provide students with a question and/or prompt and asks them to generate, rather than select, a response. Writing an essay, supplying a short answer, or completing an incomplete statement are typical constructed responses. Let's take a look now at an example of a content limit for a constructed response item in the third domain we are illustrating here -- English composition:

- English composition -- The student will be presented with a topic with which most high school students in this school would be familiar. This could be a topic dealing with a situation commonly encountered in daily living.

  The topic must embody an issue which permits the student to take one of two sides; i.e., in favor of or opposed to the proposition described.

> The prompt to the student will have three parts. One sentence will provide the student with brief background regarding the issue, with both the pro and con positions expressed in this sentence. This sentence will be labeled as Background.
>
> The background sentence will be followed by the Assignment for the student which consists of a one-sentence task description directing the student to write a paragraph in which he/she is in favor of, or opposed to the topic proposition.
>
> The assignment description will be followed by a short (no more than four sentences) paragraph giving the student sufficient detail to fully understand the assignment, expectations for the nature of the student product (e.g., take a position and support it with at least two arguments) and the nature of the criteria that will be used to judge the response.

Obviously, the content limits for the English composition assignment provide a lot of content detail about the nature of the task that is to be presented to students - the kinds of topics which are appropriate, the prompting which is to be provided, and how the assignment is to be framed. A question violating the content would need to be modified or replaced to meet the content limits.

The careful detail in the content limits, however, is necessary for several important considerations. Each student should bring a common understanding of the assignment to the task at hand, the specifications should clearly dictate these understandings. Likewise, the rater(s) of the written work should bring a common understanding of the task they are to judge; specifications help to achieve this commonality of task understanding. Further, such specification permits the generation of multiple, parallel assignments, for both instruction and for teaching, maximizing the integration of testing and instruction and the utility and fairness of the instructional process.

After the first two components of the domain description have been carefully detailed, the next task is to describe the distractor limits for selected response items, or the response criteria for constructed response items.

Distractor Limits

The distractor limits provide a description of the wrong answers or distractors that may be used as alternatives for selected response items. Based on specific categories of error types, the distractor limits define categories of wrong answers and provide rules for generating alternative responses for each item. These rules should represent common student errors and, where possible, should provide diagnostic information about the source of student error.

Here are two examples, continuing with our ongoing math and English mechanics illustrations, of distractor limits descriptions:

- Math -- distractors will be drawn from a set of shapes that are lacking in one of the following characteristics:

  3 sides
  straightness
  closed edges

- English mechanics -- distractors will be drawn from words in the sentence that are properly capitalized.

Response Criteria

Unlike selected response items, constructed response items do not include wrong answers to distract the student's choice of a correct response. In place of distractor limits descriptions, domains for constructed response items require a description of the rules and criteria that will be used to judge the quality of the student's response.

There are two judgment strategies that can be used in grading students' constructed responses -- separate criteria or holistic judgment. In the case of students' written work, for example, using the separate criteria approach might involve giving points according to how well the written product satisfies each of several distinctive criteria (such as those set forth below). On the other hand, the holistic judgment approach relies on one overall assessment of the students' work. While it is true that even in the holistic approach we use judgmental criteria to reach an overall judgment, such as the extent to which a paragraph displays acceptable organization, these criteria are applied in a comprehensive sense rather than in the criterion-by-criterion manner characteristic of the separate criteria approach.

Careful procedures need to be used during the rating process to assure reliable results. Irrespective of which judgmental approach is selected, it is imperative that those individuals who will be judging the paragraphs engage in training/clarification sessions prior to their actual judging of students' work. Judges should read the same student production, render their judgments independently, then share these judgments and discuss their reasons with other judges. Disagreements regarding the meaning of certain criteria should be resolved. This process should be continued until a high degree of inter-judge agreement is achieved (see Quellmalz & Burry, 1983, for a more detailed discussion of these issues related to writing assessment).

In addition, during the actual judging, it is desirable to have each student's work rated independently by two judges, with a third rater being called on to resolve disagreements.

Continuing with our ongoing English composition illustration, here are sample response criteria descriptions.

Organization

1. The student has written about the assigned topic.
2. The paragraph includes a topic sentence which embodies a position regarding the assigned topic.
3. All other sentences in the paragraph support the topic sentence.

Mechanics

1. The paragraph is written legibly.
2. Complete sentences are used (rather than fragment or run-on sentences).
3. Words are spelled correctly.
4. Punctuation is correct with regard to use of commas, capitalization, etc.

Now, if criteria such as the above were to be used as the basis for judging students' written essays, it would be a good idea to develop a scale with, say, one to five points, so that those judging the composition could then assign a point score to indicate the extent to which the criteria were satisfied. In addition, before using criteria such as those suggested above, it would also be a good idea to have judges make sure that they agree on what each criterion actually means: For example, that they all agree on definitions of "position," "support," and even on such mundane matters as "punctuation" since, in some instances such as the use of serial commas, correctness is as much a matter of convention as it is of hard-and-fast rules.

To this point, then, the evolving domain specification describes the domain, its content limits and, depending on the response mode in which the student will answer, either the distractor limits or the response criteria. The remaining three elements of the domain specification consist of format, directions, and sample test item.

Generating rules governing item format, directions to the student, and sample item is usually easier than generating rules for the first three parts of the specification. However, the clarity of item format and directions can have important consequences and therefore deserve careful attention. Further, parts of the specification already written may need to be modified to accommodate problems in setting up and formating the sample item.

Format

The format section of the test specifications provides a careful description of the form in which items can be presented. Again, here are some example format descriptions for the three domains we are illustrating:

- Math -- multiple choice: four shapes will be presented as response alternatives, only one of which is a proper triangle.
- English mechanics -- multiple choice: a sentence will be presented with four words or word groups from the sentence as response alternatives, one of which is incorrectly capitalized or left uncapitalized.
- English composition -- constructed: the student will be presented, aurally and written, with a three-part expository prose prompt; lined notebook paper will be provided for essay responses.

Directions

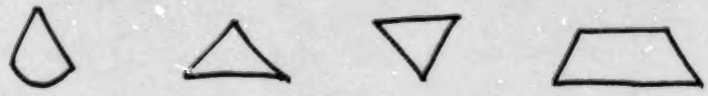
This section of the test specifications provides the actual set of directions to be used or rules for generating directions to the student for completing the test item. For example:

- Math -- Look at the four shapes below. Only one of them is a triangle. Mark an X on the shape that is a triangle.
- English mechanics -- Read the sentence below, and then circle the letter next to the word that is improperly capitalized.
- English composition -- The paragraph you write should stay on the assigned topic. In your paragraph, be sure to take a position regarding the issue and support the position you have taken. Make sure your paragraph is well organized and has appropriate grammar, spelling, and punctuation. Write clearly on the paper provided.

Sample Item

The sample item follows the rules set up in the preceding five parts of the domain specifications, and stands as a guide that test developers can follow as they develop items. Here are three examples from the domains we are using for illustration:

- Math -- Look at the four shapes below. Only one of them is a triangle. Mark an X on the shape that is a triangle.



- English mechanics -- Read the sentence below, and then circle the letter next to the word that is improperly capitalized.

  My Grandmother gave me a Timex watch for Christmas.

  A. My

  B. Grandmother

  C. Timex

  D. Christmas

° English composition --

Background: Some people think that there should be letter grades given for high school classes, while other people believe that all classes should be graded as either pass or fail.

Assignment: Write a paragraph in which you are in favor of, or opposed to, a pass/fail grading system in high school

The paragraph you write should stay on the assigned topic. In your paragraph, be sure to take a position regarding the issue and support the position you have taken. Make sure your paragraph is well organized and has appropriate grammar, spelling, and punctuation. Write clearly on the paper provided.

Summary

So far, then, we have examined and illustrated the six major features of a domain specification: The domain description provides a general statement of what's expected of the student in a particular content area; the content limits set the range of eligible content for writing test items; the distractor limits describe the wrong answers that can be used for selected response items, and the response criteria establish rules for judging a student's constructed response; the format describes the item presentation form; the directions tell the student what he or she is to do in answering the question; and the sample item, following all of the above, provides a visual aid for item writers to rely upon as they write additional items for the domain.

To show what a fully-developed domain specification might look like, we have provided an example below for 5th-grade language arts (other samples are in Appendix A).

Sample Domain Specification

Grade Level: Grade 5

Subject: Language Arts

Domain Description: Using correct capitalization in paragraphs adapted from a standard fifth grade text of a practical/informative nature.

Content Limits: The student will be presented with a paragraph of at least six sentences in which all the capital letters have been omitted. Reading level should be fifth grade or lower. The test questions will consist of identifying the words in a given sentence of the paragraph which must be capitalized. These words may include: the first word of a sentence; the names of languages, people, schools; days of the week; months of the year; places and buildings; titles of books or movies.

Each question will consist of correctly identifying all the words in one sentence, listed by number, which need to be capitalized to make the sentence correct.

Distractor Limits: The alternate responses to the questions may include: a) omission of one word(s) within the given sentence which should be capitalized; or b) listing of a word or words in the given sentence which should not be capitalized.

Format: Each sentence of the paragraph will be numbered. Each question will be multiple choice, with four possible responses.

Directions: The directions will be given: "Choose the letter which contains all the necessary capitalized words in the given sentence to make the sentence correct."

Sample Item

1. of all my high school friends, i remember jim the best. 2. he had a way of making adventures out of everyday events. 3. one sunday i remember particularly; it was a beautiful day in may. 4. i looked out the window, watching the sunlight dance on the columbia river. 5. my mom interrupted my daydreams, reminding me about my

homework for my german class. 6. i started flipping through my history book, the american republic, to avoid beginning the german grammar. 7. suddenly a hissing voice outside the window attracted my attention. 8. it was jim; he was ready for his favorite activity, fishing. 9. we sneaked down the back stairs and out the back door.

1. In the first sentence, the following words should be capitalized:

✓ a. Of, I, Jim
b. High School
c. Of
d. Of, I

These specifications, then, are the blueprint that item writers follow as they develop test items. As we will see later, the care that goes into developing domain specifications is matched by the care with which the developed items are judged to determine the extent to which they match the intentions of the domain specification. But there is one additional consideration to keep in mind while the items are being developed: The technical properties of a good item, independent of its governing domain specification. We'll take up this topic in the next section.

ITEM CONSTRUCTION RULES

When the domain specifications have been written, reviewed for substance and clarity, and checked to make sure they work as intended (e.g., the writer of the specifications can try to develop a few additional sample items, or ask a colleague to make this check to ensure there are no bugs in the specifications) item writing begins.

Items are of course written to match the specifications. But there is another consideration as well. In constructing items, there are some generally agreed upon rules, or perhaps rules-of-thumb, that help make sure that items do not contain flaws that unnecessarily cue or confuse the student. That is, having good domain specifications does not guarantee that the items generated from it will necessarily be good items. An item can have a perfect fit with its guiding specifications and yet be flawed.

In this section, then, we'll offer some rules for writing constructed response items -- essay and short answer or completion items, and for writing selected response items -- true-false, matching, and multiple choice items.

Constructed Responses

Essay items: An essay item asks the student to produce a piece of written work ranging in length from one to several paragraphs. In writing essay type items, we need to keep the following rules in mind:

1. The task expected of the student should be defined as completely as possible, without interference with the measurement of the domain being tested.
2. The topic to be written on should represent a novel situation or problem, and not be a repetition of situations or problems used for instructional purposes. If the test question merely repeats something that has happened in class, then all it requires from the examinee is recall, which is more efficiently measured by another test format, such as multiple choice.

3. To obtain adequate reliability, the student needs to have a clear picture of what constitutes an acceptable response. It is also necessary to have a detailed scoring guide. Reliability is also increased by having each student answer several questions, or by having several independent scorers per answer.
4. If students are allowed to select among several questions, each question should be of the same difficulty level.

Short answer and completion items: Each of these item forms is answered by a word, phrase, number, or other symbol that is written by the student. The two forms are essentially the same, and differ only in how the problem is presented to the student. The short answer item asks a direct question of the student, while a completion item consists of an incomplete statement to be completed by the student. Here are some rules for this item genre:

1. The question itself should not provide any extraneous clues to the answer.
2. The question must be stated so that only one brief answer is possible.
3. No grammatical clues should be given, such as "a" or "an."
4. The student should receive clear directions stating the degree of precision expected and/or the unit(s) in which the answer is to be expressed.
5. The scoring or answer key should anticipate possible synonyms or acceptable variants of the desired response.

6. Only key words should be left blank; it is generally better to have the blank (to be filled in) at the end of the statement because the student will then have in mind all information needed to give an answer (this tactic may also simplify the scoring); the blanks should be uniform in length throughout the test; where they are positioned should generally be the same.
7. Statements in which the blank would complete an instructional cliche should be avoided.

Selected Responses

True-false: The true-false item consists of a statement that the student will mark true or false, right or wrong, correct or incorrect, yes or no, fact or opinion, agree or disagree, and so forth. In each case, there are only two possible answers. Here are some rules for true-false items:

1. The item must be free from ambiguity. It must be unequivocally true or false. It is difficult to develop an item to have this property while at the same time assuring that it remains unclear to the novice student and unambiguous to the knowledgeable.
2. The question must embody only one idea.
3. The question should be stated in positive form whenever possible. It must never contain a double negative and, if it is stated in the negative form, the negative word should be clearly marked.
4. The question should be worded so that the student with only superficial knowledge would be led to the wrong answer.

5. The question can be worded so that the incorrect answer is consistent with a popular misconception or belief. It can also use phrases in false statements to give them a "ring of truth." These devices, however, must be handled carefully, since the intention should not be to trick the student.
6. The question should not depend for its truth or falsity on an insignificant word or phrase.
7. The question should not include indefinite terms, degrees, or amounts.
8. The question should not include specific determiners. True items should not be qualified; false items should not be absolute.
9. The questions should be evenly divided between true and false to help avoid biases due to guessing.
10. Material emphasized in the question should not be based on an instructional cliche.

Other Considerations: Since there are only two response options open to the student, there is a 50-50 chance of any guess being correct. True-false items are also open to criticism that the ability to recognize an incorrect statement is not necessarily dependent on knowledge of the correct answer. In general, any question that can be presented in a true-false format can usually be presented more effectively in a multiple choice format.

Matching items: A matching question consists of two columns with each word, number, or symbol in one column matched to a word, sentence, or phrase in the other. The student's job is to identify the pairs of items on the basis indicated. Here are some rules for this item form:

1. Students should be provided with clear directions explaining the basis of matching.
2. The entire item should appear on the same page.
3. Components should be short in phrasing and few in number.
4. The two columns should have appropriate labels.
5. Each alternative must be a plausible solution to all problems presented.
6. The lists should have an unequal number of components to be matched.
7. Components in the response column should be placed in some logical order.

Other considerations: It is often a good tactic to inform the student that each of the possible answers in the response column may be used more than once, just once, or not at all. This tactic, in conjunction with the provision of more responses than items to be matched, minimizes the role of guessing. For example, with one-to-one ratio, if a student knows all but two of the correct matches, he/she must choose where to place the two remaining responses. If she/he guesses correctly with either, then she/he guesses correctly with both.

The alternatives should all be of the same class of response; e.g., historical events, cities, etc. The student should not be able to discard any alternative because it is illogical, does not fit with the elements of the other column, and so forth.

Multiple choice items: A multiple choice item presents the student with a problem and a list of suggested solutions. The student is typically requested to read the stem and to select the one correct, or best alternative. The following rules apply to multiple choice items:

1. The stem should contain a complete statement of the problem to be solved.
2. The stem should be stated in clear, precise lanaguage.
3. The alternatives should be presented in a logical order: e.g., by chronology, number series, etc.
4. In a vocabulary item, the term should be in the stem, and the definition among the alternatives.
5. Either the stem should be stated in positive form, or the negative word should appear at the end of the stem and be clearly marked for emphasis.
6. All alternatives should be consistent with the grammatical and syntactical construction of the stem.
7. All alternatives should be approximately equal in length.
8. The alternatives should make reference to the item stem and not to the correct answer.
9. All alternatives should be equally attractive or plausible to the uninformed examinee.
10. The item should not have less than four alternatives (including the correct response).
11. The position of the correct answer should be evenly divided among the response options (e.g., in a 25 item test with alternatives A,

B, C, and D for each, the correct response should be evenly distributed across all four letters, in some sort of random ordering).

12. The correct answer should not be matched by an opposite distractor unless another pair of opposites is included among the other distractors.
13. The correct answer should not contain a repetition of a word or phrase found in the stem.
14. Items using pictures as stimuli should not inadvertently provide a clue to the correct answer.
15. The stem should not take a great deal of student reading time. It should contain material common to all alternatives so as to decrease reading time.
16. The correct answer should not contain an instructional cliche.

Other considerations: Items should be independent of each other; the correct answer to one question should not be necessary to obtain the correct answer to another question. Similarly, the information in the stem of one item should not become an aid in detecting the answer in another item.

"All of the above" should generally be avoided since this alternative increases the probability of the student guessing correctly by using partial information. "None of the above" should also be used with caution, since recognizing incorrect answers does not ensure that the student actually knows the correct answer.

Verbal cues to the correct answer should be avoided.

Stem and alternative language level must be appropriate to the student and to the question. For example, because of the prose used in the stem, or because of tortured construction, it may be that the item actually becomes a question assessing linguistic ability or inferential reasoning rather than assessing the intended domain.

Each item must have the same number of alternatives -- no less than four.

The student should be able to derive the problem from the stem, and should not need to read the alternatives in order to discover what question is being asked.

Although the rules we have offered above are generally accepted among test developers, some are more a matter of taste or convention than others. At any rate, as with other aspects of good test development, these rules should be provided to the people who will have the job of developing test items. They can then discuss any possible areas of disagreement and modify the rules if such modification will not jeopardize technical quality. The key notion here is that all item writers understand, accept, and apply a uniform set of rules so that when they follow the domain specifications to write their items the items will also have acceptable technical quality.

Once the items have been developed, the next job is to judge the degree to which they reflect the intentions of the domain specifications and are technically adequate. We have developed an item rating scale to help with this task.

THE ITEM RATING SCALE

Background

As we have seen, domain-referenced testing limits and defines a class of behaviors, skills, or information and provides a set of specifications which are used to generate test items reflecting the instructional process. These specifications permit the integration of testing and instruction and increase the usefulness of testing. The validity of the process depends on the match between the domain, instruction, and assessment. A critical consideration, therefore, is the extent to which the items match their specification.

Even with the most careful specification and item development process, it is likely that items will vary in the degree to which they fit or belong in the domain which they are intended to assess. Most commonly the judgment about how well an item matches or measures the domain is not a clear yes/no choice but rather should reflect the complexities of test specification and item development. The Item Rating Scale (IRS) we offer, therefore, provides a range of values that are used in judging the "belongingness" of an item to a domain. It permits judgments to be made about item compatability with each of the categories in test specifications. Further, these judgments suggest areas in which an item, or its governing specifications, may need to be modified and improved.

Using the IRS

Raters use the IRS to judge the match between the test specifications and any given item along eight independent dimensions. The first six rating categories of the IRS parallel the basic structure of domain-referenced test specifications we discussed earlier. In addition, test item features of linguistic and thinking complexity are also included in

the IRS. Just as we suggested for the item construction rules and other specified criteria, raters will need to become familiar with the IRS before they use it to judge items.

The first category of the IRS concerns the item's fit with the general domain description. The second category, content limits, compares the description of eligible subject matter and item features with the test item's contents and features. The third category judges the item against distractor limits or response criteria, depending on whether the item is a selected or constructed response type. For selected response items, the specification rules for creating wrong answer alternatives are compared with the actual wrong answer choices used in the test item. For constructed response items, the prescribed criteria for evaluating the response generated by the student are compared both to those criteria used and to the suitability of the item and conditions for eliciting a judgeable response. Format and directions are the fourth and fifth categories between specifications and actual items. In these categories the concern is whether the layout of the item and the directions for completing the test conform to the test specifications. The sample item provided in each specification is the final aspect of the test specifications included in the Item Review Scale.

Linguistic complexity and thinking complexity provide a structure for getting an accurate picture of some of the more subtle sources of complexity that may affect students' performance in a way not described or desired by the test specifications. These biasing elements are important, to the degree that the specifications and resulting items are intended to provide the same measure of performance for all students in the given area.

Raters assign a whole number value that best represents their judgment of the match between the item and its specification on the particular dimension being considered. After arriving at the rating for the item on the first dimension, i.e., domain description, raters proceed, one dimension at a time, rating the item-specification match on each dimension.

When all eight categories have been individually scored, an overall rating is then calculated for the item. The final calculations are guided by a weighting system that incorporates the scores in each category.

The ratings are then interpreted in terms of the three features judged to be most critical -- content limits, distractor limits or response criteria, and thinking complexity. These interpretations carry implications for item revision or, where necessary, specification revision.

Let's take a closer look at this process, now.

The scale we developed for this process ranges from 0 to 10, with 0 indicating a poor match between item and specification, and 10 indicating a perfect match. The scale provides the following guidelines to raters for assigning number ratings in each component of the domain specification:

0,1,2 This rating range should be used for items that are completely unrelated to the specification on the dimension you are rating.

3,4,5 This rating range should be used for items that are vaguely related and/or inadequate.

6,7 This rating range should be used for items you feel would definitely require a second look and some revision, but which you feel reluctant to totally abandon.

8,9 This rating range should be used for items that you feel are good representative match-ups with the specifications, although slightly off.

10 This rating should be used for items that are beyond a doubt perfect examples of the specification.

(The ratings representing each descriptor are a function of the multiple criteria which are used to assess each domain dimension.)

Using the scale and its suggested point-assignment guidelines, item raters should refer to the following indicators of item match with each of the eight features on which items are to be judged. Depending on the degree of match, raters then assign the item a number from 0 to 10 for each of the domain specification components.

1. Domain Description

1. The test item is a good and fair representative of the subject area outlined in the domain description of the test specifications. It does not assess an obscure or unusual aspect of the domain.

2. Test item conditions are not at odds with test intentions. This is especially important in constructed items.

3. The test item content is closely related to the instructional objective(s) stated or implied in the domain description.

2. Content Limits -- Selected Response Items Only

1. The item and additional accompanying material (e.g., graphs, maps, reading selections) follow the content limits on length and general difficulty level.

2. The item and additional accompanying material follow the content limits on eligible content, descriptive detail, and completeness of information provided.

3. The solution processes required by the student to answer the item match those described or implied in the content limits.

2. Content Limits -- Constructed Response Items Only

1. The item matches the content limits on eligible content, descriptive detail, or completeness of the prompting information provided.

2. The item provides a context for responding that is similar to that described in the content limits (e.g., time restrictions, length of written/oral response, equipment or aid restrictions, warmup or false start provisions).

3. The mental processes required by the student to respond to the item seem to match those described or implied in the content limits.

3. Distractor Limits -- Selected Response Items Only

1. The alternative answers, or distractors, provided in the item require the student to discriminate important features or factors described in the distractor limits as differentiating correct from incorrect answers. Distinctions between correct and incorrect answers are not based on trivial or irrelevant features.

2. The distractors provided in the item correspond to the content limits on number, length, and general level of difficulty.

3. Response Criteria -- Constructed Response Items Only

1. The rules used to judge the student's response are those described by the response criteria.

2. The item prompt provides a context for responding that is appropriate to the response criteria for judging the content and style/form of the response (i.e., likely to elicit a judgeable response).

3. Problems arising from incomplete or inadequate answers are dealt with in a way that upholds the testing intentions of the specifications.

4. Format

1. The organization and display (layout) of the item conform to the format description in the test specifications.

2. For selected response items only: The organization and display of any additional information (e.g., maps, graphs, pictures, reading selections) conforms to the format description.

3. For constructed response items only: The context or conditions for responding to the item (e.g., time limits, space limits, available equipment) conform to the format description.

5. Directions

1. The directions for completing the test item

correspond to the description of test directions in the test specifications.

2. The reading level and complexity of the directions follow the description of test directions in the test specifications; or seem to be within suitable range for the intended students.

6. Sample Item

1. The sample item and the test item being rated could come from the same set of items described by the test specifications.

2. The sample item and the test item are very similar in content and either distractors or response criteria.

3. The sample item and the test item are very similar in format and directions.

7. Linguistic Complexity

1. Vocabulary used in the item is consistent with the test specifications for item difficulty. Words are not used that have different or unfamiliar meanings for different students or student groups.

2. Item language structure, (including, e.g., the use of compound, complex sentences, antecedents) is consistent with the test specifications for item difficulty.

8. Thinking Complexity

1. Those mental processes required for the solution or performance of the test item, but that are not described in the domain description or content limits (i.e., are assumed), are readily available to all students at some necessary level of competence (e.g., drawing ability, handwriting legibility, short-term memory capacity, imagination, ability to separate relevant from irrelevant, detail from generalization).

2. Directions for completing the test item provide the same amount of information and structure for all students. Everyone has the same understanding of what is expected and of what the limits or rules for answering are.

3. For items with nonverbal components, it is reasonable to assume that these components comform with the content limits or distractor limits in their intended meaning, and that this interpretation is stable across all groups of test takers.

Each item is rated on each of the above eight components, and the separate ratings it received are recorded. It is also helpful if raters provide written comments explaining why a particular rating was assigned. This kind of information is extremely useful in situations where feedback is needed for revising items or specifications.

Overall Item Rating

We suggest that an item's overall rating be interpreted primarily, but not exclusively, against the three most critical features of the domain specifications -- content limits, distractor limits or response criteria, and thinking complexity. We suggest this not only because these three features provide the substantive core of any test item, but also because problems in these areas are more difficult to correct than problems in other components of the domain specification.

The overall item rating process reflects the importance of these three critical elements. The original ratings the item received on the three critical features are first weighed, or multipled, by a factor of three, and then summed with the ratings from the other specification dimensions.

In this way, an item can receive a final weighted score ranging from 0 to 140, reflecting a highest possible rating of 10 each on five of the

domain components, and a highest possible rating of 30 (3 X 10) on each of the three critical elements.

The total score an item receives through this process is then divided by 14 (i.e., the total number of points on the weighted scale -- five of the elements are unweighted and each has a value of 1; three of the components are weighted and each has a value of 3) and the quotient, ranging from 0 to 10, is the item's overall rating.

After or during the rating process, items will also need to be reviewed for their adherence to technical rules of item construction listed earlier. Where problems occur, items will need to be revised accordingly, if the item is judged to be an adequate match to its specification. Should recurring technical flaws of the same type be found, revisions or additions to the domain specifications may be necessary.

Let's take a moment now to consider the issue of item-rating interpretation.

Interpreting an Item's Overall Rating

Because we are concerned about an item's validity and the accuracy with which it assesses student knowledge in the given domain, we suggest that high quality standards be applied to the interpretation process. That is, with a possible final rating of 0 to 10, we suggest that the first criterion to consider is whether or not an item receives an overall rating of at least 7 or 8 points.

After this determination has been made, we can then take a much closer look at items with an overall rating of 7 or better and those with an overall rating below 7 points. The point of this process is to ascertain how much the critical, weighted features contribute to an item's overall rating, and to base our decisions accordingly about keeping an item as is, modifying an item, rejecting an item, or revising the original domain specifications.

We offer the following suggestions to guide the interpretation process.

Items Rated 7 or Better:

(1) If all three of the weighted, critical features are rated 8 or better (here and elsewhere this statement means before rating weights were applied), then the item is good and basically in comformity with the test specification. Any further item review and/or rewrite process should be directed toward other domain features on which the item received a lower rating.

(2) If any one of the critical features received a rating of 7 or lower, it will be necessary to return to the original item specifications guiding that feature, so as to better align the item with the described testing intentions. An item in this category likely has problems in other, non-weighted features. It will probably be necessary to rewrite the item and examine the revision to make sure that all critical features are still up to par.

(3) If more than one critical feature received a rating of 7 or lower, the item has serious validity problems. If it is decided that this item

exemplifies the kind of test item that is actually desired, then it will be necessary to reconsider its guiding specifications. They may need to be better conceptualized, reconceptualized, or become more complete in their descriptions of desired item qualities. If, however, the specifications, as written, are close to the actual testing intentions, then the item should be discarded from the pool and replaced with a new one.

The statements provided previously in the IRS rating categories can be used as a guide in item review and revision.

Items Rated Below 7:

(1) If all three critical, weighted features are rated 8 or better, the item is potentially a good one but it has serious problems in presentation. It will again be necessary to return to the original specifications guiding those features receiving low ratings, and to revise the item's manner of presentation.

(2) If one or more of the critical features scored 7 or lower, the item probably is not worth any attempted fix-up effort. However, before starting the item writing process again, it would be a good idea to reconsider the guiding specifications; they may need to be better conceptualized or provide fuller descriptions of the desired item features.

Again, the IRS statements describing item match can be used in item review and revision.

USING THE GUIDE

Appendix C contains copies of the materials that a district or school will need if they wish to apply the item review process we have described here. The appendix provides the directions to raters, an individual rating

worksheet, an overall item rating form, and the guide to interpreting overall ratings.

Should a district or school develop domain specifications following the procedures in this guide, and then have item writers develop items to measure these domains, the materials in Appendix C can be reproduced to guide the local item review process. Raters would then be given a copy of the domain specifications, the items written for it, and the materials we offer to facilitate the rating process.

To guide the item development process, the district or school should also familiarize item writers with the construction rules we offered earlier in the guide. To guide the item rating process, the district or school should provide raters with a copy of the indicators of item-domain match that also appeared earlier in the guide.

For districts or schools who wish to implement the procedures in the guide, we strongly suggest that a local staff member (or outside expert) be designated as facilitator. That person should become thoroughly familiar with the contents of the guide, and take responsibility for leading discussion/orientation sessions on domain specifications, item writing, and item rating and interpretation with the staff members who will be responsible for carrying out these activities.

In addition, the local facilitator should take responsibility for making any materials adaptations that are appropriate in the local setting, and then oversee the entire process of domain specification, item development, and item review.

CONCLUSION

In this guide we have provided a means of developing domain specifications which create a direct link between instruction and assessment. These specifications establish rules which can be applied to develop items and tests which represent the domain of instructional concern and provide accurate assessment of student learning in that domain.

While this match between an item and its specification is of primary importance, it is also important to assure that test items are free from technical flaws. To maximize this possibility we have outlined some generally accepted rules of item construction.

In addition, no matter how clear the domain specification or how skilled the item writer, some items will provide a better fit with a given domain than others. The item review scale we have described offers a way to systematically judge an item's fit and to obtain information suggesting areas in which an item, or its domain specification, need to be improved.

We admit here that, initially at least, such a painstaking approach to test planning, development, and review requires some time. We can point out, however, that people who have attended the workshops we have conducted on the topics in this guide report that after practice the process tends to become internalized and developing tests in such careful fashion becomes routine.

Finally, given the increasing need to develop tests which are appropriate to local needs and sensitive to local instructional practice and intentions, we believe that the initial investment of time is well worth the effort.

REFERENCES

Burry, J., Dorr-Bremme, D.W., Herman, J.L., Lazar-Morrison, C.M., Lehman, J.D., & Yeh, J.P. Teaching and testing: Allies or adversaries? CSE Report No. 165. Los Angeles: UCLA Center for the Study of Evaluation, 1981.

Burry, J., Catterall, J., Choppin, B., & Dorr-Bremme, D.W. Testing in the nation's schools and districts: How much? What kinds? To what ends? At that costs? CSE Report No. 194. Los Angeles: UCLA Center for the Study of Evaluation, 1982.

Dorr-Bremme, D.W. Assessing students: Teacher's routine practices and reasoning. Evaluation Comment, 1983, 6(4), 1-12.

O'Shea, D.W. School district evaluation units: Problems and possibilities. In A. Bank & R.C. Williams, (Eds.) Evaluation in school districts: Organizational perspectives. CSE Monograph No. 10. Los Angeles: UCLA Center for the Study of Evaluation, 1981.

Quellmalz, E.S., & Burry, J. Analytic scales for assessing students' expository and narrative writing. CSE Resourse Paper No. 5. Los Angeles: UCLA Center for the Study of Evaluation, 1983.

APPENDIX A

SAMPLE DOMAIN-REFERENCED TEST SPECIFICATIONS

Grade Level: Grade 9

Subject: English-punctuation

Domain Description Correctly punctuating given paragraphs adapted from a standard eighth grade text of a practical/informative nature.

Content Limits: The student will be presented with one paragraph in which all the correct punctuation marks have been omitted, except for apostrophes in contractions (I'll) and possessives (Jane's), dashes, and semi-colons.

For each question, students will be asked to choose all the correct punctuation marks which must be added in a given sentence to make the sentence correct. The punctuation marks to be identified and added may include:

a. periods at the end of a declarative or imperative sentence, after an abbreviation, or an initial
b. question marks following an interrogative sentence
c. exclamation point after exclamatory sentences or interjections
d. colon after the salutation in a business letter, or to separate minutes and hours in expressions of time, and to show that a series of things or events follows
e. quotation marks enclosing a quotation or a fragment of it, enclosing the title of a story or poem which is a part of a larger book
f. comma in a date or address; to set off such words as "yes" at the beginning of a sentence; to set off names of persons or words (phrases) in apposition; to separate words in a series, direct quotations, parallel adjectives, parenthetical phrases; after introductory prepositional phrases; before coordinate conjunctions; after the salutation and closing in a friendly letter; to separate a dependent clause from an independent clause in a complex sentence.

Distractor Limits The alternate responses to the questions may include:

a. omission of punctuation mark (s) within a given sentence which should be included, or
b. inclusion of a punctuation mark or marks which is not necessary or correct in the given sentence

Directions: The directions will be given: "Choose the letter which contains all the necessary punctuation marks in the given sentence which will make the sentence correct."

Format: Each question will be multiple choice, with four possible responses.

Sample Item: 1. If she starts to sing again I'll crack up 2. It is funny how it hurts to hold back a laugh 3. I was sitting in the auditorium at 10:00 am and we were having a singing rehearsal for graduation 4. Sit up Get off those shoulders Think tall Sing tall Sing like this said Ms Small 5. I knew that if she was going to tweet like a bird again I would laugh 6. But I just could not laugh because Ms Small would kick me out of the auditorium and that meant Felson's office--and no graduation 7. La la la--sing children Sing with your hearts said Ms Small 8. I couldn't hold it 9. She was so funny I almost rolled off the auditorium seat 10. The other students didn't laugh but I sounded like Santa Claus 11. It became quiet for a second 12. What are you doing Joe I know it is you Present yourself to Mr Felson at once that voice said 13. Ms Small is a foot shorter than a tall Coke but she has the bark of a hungry hound dog

1. The first sentence should be written:

a. If she starts to sing again I'll crack up.
b. If she, starts to sing again, I'll crack up
✓ c. If she starts to sing again, I'll crack up.
d. if she starts, to sing again, I'll crack up.

Grade Level: Grade 8

Subject: Introduction to Algebra

Domain Description: Using basic operations and laws governing open sentences, solve equations with one unknown quantity.

Content Limits:

1. Stimuli include a number sentence with one unknown quantity, represented by a lower case letter in italics, and an array of four solution sets or single answers, only one of which is correct.
2. Number sentences may be statements of equalities or inequalities.
3. The number sentences may require simplifying before solving by combining like terms or carrying out operations indicated (e.g., by parentheses).
4. Number sentences will have no more than five terms. Fractions may be used but not decimal fractions and non-decimal fractions in the same expression. Exponents (powers) may appear in the expression only if they cancel out and need not be solved or modified.
5. Solution sets for equations and inequalities will be drawn from the set of rational numbers ($\pm$). The null set ($\emptyset$) may be used as a correct solution set.
6. Factoring may be a requisite operation for solving the equation.
7. Application of the distributive property of multiplication and the use of reciprocal values may be requisite operations for solving the equation.

Distractor Limits:

1. Distractors may be drawn from the set of wrong anwers resulting from errors involving any one of the following operations:
   a. combining terms
   b. transformations that produce equivalent equations (e.g., transferring terms using the principle of reciprocal values)
   c. distributing multiplication, with positive or negative numbers (e.g., across parentheses)
   d. carrying out basic operations using brackets or parentheses

2. Distractors may also be drawn from the set of wrong answers due to incomplete solution sets.

3. Distractors may not reflect errors due to wild guessing, calculations involving negative numbers, errors in basic operations.

4. "None of the above" is not an acceptable alternative.

Format: Multiple choice; five alternatives.

Directions: Solve the equation. Then select the correct answer or solution set from the choices given.

Sample Item:

1. $8n + 2 = 2n + 38$; $n = ?$

a) $n = 3$
✓ b) $n = 6$
c) $n = 4$
d) $n = 5$
e) $n = 7.6$

2. $16x \leq 32$; $x = ?$

a) $x = 48$
✓ b) $x = \{0,1,2\}$
c) $x = 2$
d) $x = \emptyset$
e) $x = \{3,4,5...\}$

Grade Level: Secondary

Subject: Life science - circulatory system

Domain Description: Applying understanding of the circulation system to predict cause-effect relationships within the system.

Content Limits:

1. Circulatory systems include: pulmonary circulation, coronary circulation, systemic circulation (renal and portal).

   Heart structures eligible for indentification and differentiation of function include: left and right atria (or auricles), left and right ventricles, pulmonary artery and veins, systemic artery and veins, aorta, valves.

   Other structures eligible: veins, arteries, capillaries, femoral artery and vein, inferior vena cava and superior vena cava, jugular vein and carotid artery, brachial artery and basilic vein, portal and renal veins and arteries.

   Eligible cause-effect situations include: heart attack, arteriosclerosis, injury to aorta or other major veins and arteries (superior, inferior vena cava, jugular, carotid, femoral veins/arteries, portal and renal veins and arteries, brachial and basilic), high blood pressure, pulse, heart murmur.

2. Items on cause effect may present the cause and ask the effect or vice versa. These items may be presented pictorially, e.g., showing blood clot in the coronary artery. However, in these cases, all parts must be labelled for the student.

Response Criteria:

1. For labelling pictures, terms must be correct; spelling does not count. Partial credit may be given for correct labels in pictures requiring more than one response; incorrect labelling that affects meaning (e.g., not including the word artery or vein in carotid), should be counted as incorrect.

2. Correct responses to cause-effect must include all the underlined points below in order to be considered a complete and correct answer. Partial credit may be awarded at the discretion of the teacher.

   a. heart attack: clot in coronary artery preventing the flow of blood to the heart; heart tissue damaged or destroyed due to lack of food and oxygen since blood can't reach cells.

   b. injury to major veins and arteries: should differentiate the functions and locations of the given vein or artery (femoral artery and vein; inferior and superior vena cava; jugular vein and carotid artery; brachial artery and basilic vein; portal and renal veins and arteries; aorta).

   c. arteriosclerosis: loss of elasticity of artery walls which normally stretch and relax with the pulsing during heartbeat. Lost elasticity, often due to fatty deposits on the artery walls (hardening of the arteries), can create abnormally high blood pressure as the blood is pushed through narrower ducts.

   d. high blood pressure: could describe two possible causes--exercise (heart pumps harder to supply more oxygen to the muscles), and changes to the blood vessels, e.g., arteriosclerosis (smaller tube way for blood flow increases pressure).

   e. pulse and heartbeat: should describe the pumping action of the heart as reflected in the arteries, stretching the arterial walls, pulse as accurate indicator of heart action.

   f. heart murmur: must describe valve functions, normally and their sound (ventricles contract and valves close; ventricles relax and aorta valves open). Murmur represents backflow of blood from incomplete or improper valve closing.

Format: Fill-in; label figures; or paragraph responses.

Directions: Complete each sentence. OR Label each part of the diagram representing _____. OR Diagram (or describe) the _____ process through the heart. OR Answer each question completely, including a description of causes, effects, and other processes involved.

Sample Item: Answer completely, including a description of parts or functions where necessary.

What would be the effect of injury to the carotid artery?

APPENDIX B

ITEM DIFFICULTY LEVELS

AN ANNOTATED COGNITIVE DOMAIN TAXONOMY*

This classification describes, from simplest to most complex, six degrees to which information that is taught can be learned.

1. Knowledge. Recalling information pretty much as it was learned. In its simplest manifestation, this includes knowledge of the terminology and specific facts-dates, people, etc., associated with an area of subject matter. At a more complex level it means knowing the major sub-areas, methods of inquiry, classifications, and ways of thinking characteristic of the subject area, as well as its central theories and principles.

2. Comprehension. Reporting information in a way other than how it was learned in order to show that it has been understood. Most basically this means reporting something learned through an alternative medium. More complex evidence of comprehension involves interpreting information in "one's own words" or in some other original way, or extrapolating from it to new but related ideas and implications.

3. Application. Use of learned information to solve a problem. This means carrying over knowledge of facts or methods learned in one specific context to completely new ones.

4. Analysis. Taking learned information apart. Analysis means figuring out a subject matter's most elemental ideas and their interrelationships.

5. Synthesis. Creating something new and good, based on some criterion. This creation can be something that communicates to an audience, that plans a successful goal-directed endeavor, or that subsumes a collection of ideas within a new theory.

6. Evaluation. Judging the value of something for a particular purpose. This means making a statement of something's worth based either on one's own well-developed criteria or on the well-understood criteria of another.

---

* Adapted from TAXONOMY OF EDUCATIONAL OBJECTIVES: The Classification of Educational Goals: HANDBOOK I: Cognitive Doman, by Benjamin S. Bloom, et al. Copyright 1956 by Longman Inc. Previously published by David McKay Company, Inc. By permission of Longman Inc.

APPENDIX C

MATERIALS USED IN ITEM RATING PROCESS

DIRECTIONS

The Item Rating Scale (IRS) is intended for use in making systematic content validity judgments for domain-referenced tests by comparing test specifications with items. The scale is devised in such a way as to provide feedback, as well, for revising items or specifications as necessary. In using the IRS, one test item at a time is rated against a set of test specifications.

1. Get a copy of the test specifications and the items you wish to rate.

2. Go through the categories of the IRS using the statements in each section to direct you in judging the compatability of your item with the six test specification features and the two additional categories concerned with complexity issues.

3. In each section, rate the extent to which your item appears to be a member of the hypothetical set of items described by the test specifications in that category. Use a scale of 0 to 10 to rate your item, letting 0 indicate a poor match and 10 a perfect one.

   The following guidelines are suggested for assigning number ratings in each section:

0,1,2	This rating range should be used for items that are completely unrelated to the specification in the dimension you are rating.
3,4,5	This rating range should be used for items that are vaguely related and/or inadequate.
6,7	This rating range should be used for items you feel would definitely require a second look and some revision, but which you feel reluctant to totally abandon.
8,9	This rating range should be used for items that you feel are good representative match-ups with the specifications although slightly off.
10	This rating should be used for items that are beyond a doubt perfect examples of the specifications.

   Enter your rating in the box provided.

   Space for taking notes has been provided. It is strongly suggested that you take advantage of this to make comments about the item as you rate it. Such notes will be useful later in revising the item or the specifications.

4. Complete the Overall Item Rating sheet by carrying over the rating scores from each section to the appropriate line of the rating sheet. Make the calculations indicated in the directions there, applying the rating weights where indicated.

5. Refer to the Interpretation Guide for rating explanations.

6. REMEMBER YOU ARE RATING THE MATCH BETWEEN THE ITEM AND THE SPECIFICATION, NOT THE ITEM AND YOUR EXPECTATIONS OR STANDARDS! ALSO, EACH IRS CATEGORY SHOULD BE RATED INDEPENDENTLY OF THE OTHERS; FOR EXAMPLE, DOMAIN DESCRIPTION RATINGS DO NOT INCLUDE CONTENT LIMIT CONSIDERATIONS. USE THE STATEMENTS PROVIDED TO GUIDE YOUR JUDGMENTS.

SPECIFICATION BEING RATED _____

RATER TITLE _____

COMMENTS: (additional comments can be made on the reverse side)

ITEM NUMBER				
Domain Description				
*Content Limits				
*Distractor Limits or Response Criteria				
Format				
Directions				
Sample Item				
Linguistic Simplicity				
*Thinking Complexity				
TOTAL				
÷ 14 =				

SPECIFICATION BEING RATED _____

RATER TITLE _____

COMMENTS: (additional comments can be made on the reverse side)

ITEM NUMBER

Domain Description			
*Content Limits			
*Distractor Limits or Response Criteria			
Format			
Directions			
Sample Item			
Linguistic Simplicity			
*Thinking Complexity			
TOTAL			
÷ 14 =			

55

OVERALL ITEM RATING

1. Recopy item ratings from each section, making the indicated weighting adjustments for the starred features: Content Limits, Distractor Limits or Response Criteria, and Thinking Complexity.

DOMAIN DESCRIPTION		____
*CONTENT LIMITS	(____x 3)	= ____
*DISTRACTOR DOMAIN OR RESPONSE CRITERIA	(____x 3)	= ____
FORMAT		____
DIRECTIONS		____
SAMPLE ITEM		____
LINGUISTIC COMPLEXITY		____
*THINKING COMPLEXITY	(____x 3)	= ____
	TOTAL	____

2. Total the scores. Divide the total by 14. This number is the overall item rating.

OVERALL ITEM RATING ____ ÷14 = ____

3. Item's technical adequacy:

Acceptable _____

Modifications needed _____

4. Refer to the Interpretation Guide for assistance in making decisions about the item and for suggestions on changing the item or its specifications.

INTERPRETATION GUIDE

ITEMS RATED 7 OR BETTER

IF ALL THREE STARRED CRITICAL FEATURES ARE RATED 8 OR BETTER*, your item is good, basically in conformity with the test specifications. Review and rewrite efforts should be directed toward other features that scored low, e.g., Format. Use the statements in the IRS rating categories to guide your work.

IF ONE CRITICAL FEATURE RECEIVED A RATING OF 7 OR LOWER*, go back to the specifications on that feature. Try to better align your item with the testing intentions described in the specifications. Use the statements in the IRS to help direct your thinking. You may also have problems with other features. Rewrite the item but review it again to be certain all critical features are up to par.

IF MORE THAN ONE CRITICAL FEATURE RECEIVED A RATING OF 7 OR LOWER*, the item has serious validity problems. If this is the kind of test item you want, then you should reconsider the specifications you are using. They may need to be better conceptualized, reconceptualized, or more complete in their description of item qualities. If the specifications are closer to what you want to be testing, throw out the item. Find or write a new item.

ITEMS RATED BELOW 7

IF ALL THREE STARRED CRITICAL FEATURES ARE RATED 8 OR BETTER*, your item is potentially a good one but has serious problems in presentation. Go back to the specifications for those features receiving the low ratings. Clean up your item. Use the statements in the IRS rating categories to guide your efforts.

IF ONE OR MORE OF THE CRITICAL FEATURES SCORED 7 OR LOWER*, your item isn't worth the fix-up effort. Before you start over, reconsider the specifications with which you are working; they may need to be better conceptualized or more complete in their description of item features.

* Before rating weights are applied.